Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 492–508

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Enriching machine-mediated speech-to-speech translation using contextual information☆

Vivek Kumar Rangarajan Sridhar[a,*], Srinivas Bangalore[a], Shrikanth Narayanan[b]

[a] *AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932, United States*

[b] *University of Southern California, Ming Hsieh Department of Electrical Engineering, 3740 McClintock Avenue, Room EEB430, Los Angeles, CA 90089 2564, United States*

Received 10 April 2010; received in revised form 27 July 2011; accepted 30 August 2011
Available online 21 September 2011

Abstract

Conventional approaches to speech-to-speech (S2S) translation typically ignore key contextual information such as prosody, emphasis, discourse state in the translation process. Capturing and exploiting such contextual information is especially important in machine-mediated S2S translation as it can serve as a complementary knowledge source that can potentially aid the end users in improved understanding and disambiguation. In this work, we present a general framework for integrating rich contextual information in S2S translation. We present novel methodologies for integrating source side context in the form of dialog act (DA) tags, and target side context using prosodic word prominence. We demonstrate the integration of the DA tags in two different statistical translation frameworks, phrase-based translation and a bag-of-words lexical choice model. In addition to producing interpretable DA annotated target language translations, we also obtain significant improvements in terms of automatic evaluation metrics such as lexical selection accuracy and BLEU score. Our experiments also indicate that finer representation of dialog information such as *yes–no questions*, *wh-questions* and *open questions* are the most useful in improving translation quality. For target side enrichment, we employ factored translation models to integrate the assignment and transfer of prosodic word prominence (pitch accents) during translation. The factored translation models provide significant improvement in assignment of correct pitch accents to the target words in comparison with a post-processing approach. Our framework is suitable for integrating any word or utterance level contextual information that can be reliably detected (recognized) from speech and/or text.

Keywords: Speech-to-speech translation; Rich context; Dialog acts; Prosody

1. Introduction

Conventional approaches to speech-to-speech (S2S) translation (Gao et al., 2006; Nakamura et al., 2006; Stallard et al., 2007) typically compartmentalize the task into three independent subtasks. The source speech is transcribed into source text using automatic speech recognition (ASR), followed by translation of the source text into target text using machine translation (MT), and finally, the target text is synthesized into target speech using text-to-speech (TTS)

☆ This paper has been recommended for acceptance by 'Guest Editors Speech–Speech Translation'.

* Corresponding author. Tel.: +1 973 360 7018.
E-mail addresses: vkumar@research.att.com (V.K. Rangarajan Sridhar), srini@research.att.com (S. Bangalore), shri@sipi.usc.edu (S. Narayanan).

0885-2308/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2011.08.001

synthesis. The predominant focus of technology development has been on optimizing individual component metrics such as word error rate for ASR, BLEU score (Papineni et al., 2002) for MT and subjective listening tests for TTS. The serial processing paradigm can potentially propagate errors and exacerbate the uncertainty in the output hypotheses as each of the blocks are inherently imperfect. While we have made substantial progress in integrated optimization of speech recognition and machine translation (Matusov et al., 2005; Bertoldi et al., 2008), the focus has always been on maximizing the *content* (as conveyed through the words). On the other hand, augmenting the conventional approach with contextual information such as discourse structure, word prominence, emphasis, and contrast can play a vital role in improving the overall quality of communication. While previous efforts have investigated the use of dialog act tags and prosodic information in S2S translation for dialog modeling and linguistic analysis (Reithinger et al., 1996; Reithinger and Engel, 2000; Nöth et al., 2000), we are interested in exploiting such contextual information in machine-mediated cross-lingual communication using S2S translation systems. The ultimate objective of machine-mediated S2S translation is the accurate transfer of conceptual information to facilitate a dialog. While maximizing objective metrics can improve the individual components of S2S translation, focusing only on the content may be sub-optimal in terms of overall communication accuracy. We believe that accurate recognition of contextual information can benefit state-of-the-art in machine-mediated S2S translation in the following ways.

- Augment the output hypothesis with rich information that can aid an end user in understanding and disambiguation.
- Serve as additional features that can potentially improve machine translation.
- Improve the quality of text-to-speech synthesis.
- Aid in the natural flow of dialog.

In this work, we present a novel framework for incorporating contextual information in conventional speech-to-speech translation systems. Our enriched S2S translation framework can exploit contextual information from both the source side and target side of translation. Specifically, we focus on the integration of dialog act (DA) tags and prosodic word prominence in S2S translation (Rangarajan Sridhar et al., 2008a,b). First, we describe a maximum entropy classification framework for automatically tagging utterances and words with dialog act tags and prosodic prominence, respectively. Second, we describe specific techniques for integrating such contextual information in S2S translation. While the dialog act tags are automatically detected from the source language to improve machine translation accuracy, the prosodic word prominence labels are exploited on the target side to enrich the target text for improved text-to-speech synthesis. Finally, we outline some outstanding issues and challenges for exploiting rich contextual information in cross-lingual dialog based communication.

The rest of the paper is organized as follows. Section 2 summarizes previous work in automatic recognition of rich contextual information and their subsequent use in spoken language processing. In Section 3, we present the formulation for the proposed enriched speech-to-speech translation framework. In Section 4, we describe the maximum entropy models used for tagging units (words or utterances) with dialog act tags and prominence, and present details about the data used in this work in Section 5.

2. Related work

Enriched transcription has emerged as a unifying theme in spoken language processing combining automatic speech recognition, speaker identification and natural language processing with the goal of producing richly annotated speech transcriptions that are useful both to human readers and to automated programs for indexing, retrieval and analysis. For example, these include punctuation detection (Byron et al., 2002), topic segmentation, disfluency detection and clean-up (Liu et al., 2003), semantic annotation, dialog act tagging (Rangarajan Sridhar et al., 2009), pitch accent and boundary tone detection (Wightman and Ostendorf, 1994; Rangarajan Sridhar et al., 2008c), as well as speaker segmentation, recognition, and annotation of speaker attributes. These meta-level tags can be considered to be an intermediate representation of the context of the utterance along with the content provided by the orthography. It is only but a natural step to exploit such meta-level tags in speech-to-speech translation.

Recently, machine translation has focused on incorporating syntactic and morphological information (Yamada and Knight, 2001; Koehn and Hoang, 2007; Hassan et al., 2007; Avaramidis and Koehn, 2008) to augment the translation using words and phrases. However, the use of such meta-level information is typically motivated by potential improvements in objective metrics such as BLEU score. They are not used from the perspective of improving understanding

or providing assistance in downstream processing, essential in S2S setting. On the other hand, several interlingua translation schemes have used the notion of semantic concepts to model the translation process (Bennett, 1989; Dorr, 1992; Mayfield et al., 1995; Levin et al., 1998). While the use of semantic concepts is attractive for promoting better understanding of source sentences, it involves manual design and typically works only for limited domains. Exploiting rich information (semantic concepts, dialog act tags, prosody) in statistical speech translation can potentially coalesce the strengths of interlingua and corpus-based approaches.

Dialog act tags have been previously used in the VERBMOBIL statistical speech-to-speech translation system (Reithinger et al., 1996). In that work, the predicted DA tags were mainly used to improve speech recognition, semantic evaluation, and information extraction modules. A dialog act based translation module in VERBMOBIL was presented in Reithinger and Engel (2000). The module was mainly designed to provide robustness in the translation process in case of defective input from the speech recognition system. Ney et al. (2000) proposed a statistical translation framework to facilitate the translation of spoken dialogues in the VERBMOBIL project. Their framework was integrated into the VERBMOBIL prototype system along with the dialog act based approach developed in (Reithinger and Engel, 2000). Discourse information in the form of speech acts has also been used in interlingua translation systems (Lavie et al., 1996) to map input text to semantic concepts, which are then translated to target text.

Prosodic information has mainly been used in speech translation for utterance segmentation (Matusov et al., 2007; Fügen and Kolss, 2007) and disambiguation (Nöth et al., 2000). The VERBMOBIL speech-to-speech translation (S2S) system (Nöth et al., 2000) utilized prosody through clause boundaries, accentuation and sentence mood for improving the linguistic analysis within the speech understanding component. The use of clause boundaries improved the decoding speed and disambiguation during translation. More recently Agüero et al. (2006) have proposed a framework for generating target language intonation as a function of source utterance intonation. They used an unsupervised algorithm to find intonation clusters in the source speech similar to target speech. However, such a scheme assumes some notion of prosodic isomorphism either at word or accent group level.

Dialog act tags and prosodic information have also been used in text-to-speech synthesis to improve the naturalness of the synthesized speech (Syrdal and Kim, 2008). A variety of techniques have been used for exploiting contextual information, ranging from, rule-based generation (Katae and Kimura, 1996), parametric methods of prosody prediction (Ross and Ostendorf, 1999; Raux and Black, 2003; Kei Fujii and Campbell, 2003; Sakai and Glass, 2003), prediction of symbolic labels (e.g., ToBI) and generation of *f0* contours from the predicted labels (Campbell and Black, 1996; Black and Hunt, 1996), and by using prosodic databases that facilitate unit selection of sub-word units with appropriate intonation labels (Bulyko and Ostendorf, 2001). With the availability of such diverse techniques for exploiting dialog and prosodic information in TTS, it is critical that such information is detected and transmitted for downstream processing in S2S translation.

3. Enriched speech-to-speech translation

The general problem of enriched statistical speech-to-speech translation can be summarized as follows. If S_s, T_s and S_t, T_t are the speech signals and equivalent textual transcription in the source and target language, L_s, L_t the enriched representation for the source and target language, we can formalize our proposed S2S translation as shown in Fig. 1. Eq. (2) is obtained from Eq. (1) through conditional independence assumptions. Even though the recognition and translation can be performed jointly (Matusov et al., 2005), typical S2S translation frameworks compartmentalize the ASR, MT and TTS with each component maximized for performance individually. T_s^*, T_t^* and S_t^* are the arguments maximizing the ASR, MT and TTS components, respectively. L_s^* is the rich annotation detected from the source speech signal S_s and text T_s^*. L_t is the rich annotation in the target language. In this work, we do not address the speech synthesis part and assume that we have access to the reference transcripts or 1-best recognition hypothesis of the source utterances. The rich annotations (L_s, L_t) can be prosody (Agüero et al., 2006), syntactic or semantic concepts (Gu et al., 2006), or, as in this work, dialog act tags and prosodic prominence.

4. Maximum entropy classification

We use a maximum entropy sequence tagging model for the purpose of automatic tagging. Given a sequence of linguistic $U = u_1, u_2, \cdots, u_n$ and a class vocabulary ($c_i \in \mathcal{C}, |\mathcal{C}| = K$), we need to predict the best tag sequence $C^* = c_1, c_2, \cdots, c_n$. The classifier is used to assign to each linguistic unit a label (prosodic prominence or dialog act tag)

$$S_t^* = \arg\max_{S_t} P(S_t|S_s) \quad (1)$$

$$P(S_t|S_s) = \sum_{T_t,T_s,L_s,L_t} P(S_t, T_t, T_s, L_s, L_t|S_s)$$

$$= \sum_{T_t,T_s,L_s,L_t} P(S_t|T_t, L_t, T_s, L_s, S_s).P(T_t, L_t|T_s, L_s, S_s).P(L_s|T_s, S_s).P(T_s|S_s) \quad (2)$$

$$\approx \sum_{T_t,T_s,L_s,Lt} P(S_t|T_t, L_t, L_s).P(T_t, L_t|T_s, L_s).P(L_s|T_s, S_s).P(T_s|S_s) \quad (3)$$

$$\max_{S_t} P(S_t|S_s) \approx \max_{S_t} P(S_t|T_t^*, L_t^*, L_s^*). \max_{T_t,L_t} P(T_t, L_t|T_s^*, L_s^*). \max_{L_s} P(L_s|T_s^*, S_s). \max_{T_s} P(T_s|S_s) \quad (4)$$

Augmented Text-to-Speech **Enriched Machine Translation** **Rich Annotation** **Speech Recognition**

Fig. 1. Formulation of the proposed enriched speech-to-speech translation framework. (a) Conventional speech-to-speech translation and (b) context enriched speech-to-speech translation.

conditioned on a vector of local contextual feature vectors ($\mathbf{\Phi}$) comprising the lexical, syntactic and acoustic-prosodic information. We used the machine learning toolkit LLAMA (Haffner, 2006) to estimate the conditional distribution using Maxent. Conditional random fields (CRFs) (Lafferty et al., 2001) are also a possibility for the tasks considered in this work:

$$C^* = \arg\max_C P(C|U) \quad (5)$$

$$\approx \arg\max_C \prod_{i=1}^{n} P(c_i|\mathbf{\Phi}(u_{i-k}^{i+l})) \quad (6)$$

$$= \arg\max_C \prod_{i=1}^{n} P(c_i|\mathbf{\Phi}(W, S, A, i)) \quad (7)$$

where $\mathbf{\Phi}(W, S, A, i) = (w_{i-k}^{i+k}, s_{i-k}^{i+k}, \mathbf{a}_{i-k}^{i+k})$ are a set of lexical, syntactic and acoustic features extracted within a bounded local context k. In this paper, the lexical features are word trigrams, syntactic features are trigrams of part-of-speech tags and supertags (Bangalore and Joshi, 1999) and acoustic-prosodic features are trigrams of the normalized (over utterance) f0 and energy values extracted over 10 ms frames.

The dialog act classifier was trained on the Switchboard-DAMSL corpus (Jurafsky et al., 1998). The Switchboard-DAMSL (SWBD-DAMSL) corpus consists of 1155 dialogs and 218,898 utterances from the Switchboard corpus of telephone conversations, tagged with discourse labels from a shallow discourse tagset. The original tagset of 375 unique tags was clustered to obtain 42 dialog tags as in (Jurafsky et al., 1998). In addition, we also grouped the 42 tags into 7 disjoint classes, based on the frequency of the classes and grouped the remaining classes into an "Other" category constituting less than 3% of the entire data. The simplified tagset consisted of the following classes: *statement, acknowledgment, abandoned, agreement, question, appreciation, other*.

The prosodic prominence classifier was trained on a subset of the Switchboard corpus that had been hand-labeled with pitch accent markers (Ostendorf et al., 2001). The corpus is based on about 4.7 h of hand-labeled conversational speech (excluding silence and noise) from 63 conversations of the Switchboard corpus and 1 conversation from the CallHome corpus. The corpus contains about 67k word instances (excluding silences and noise). Prominent syllables were marked only with "*" for indicating a pitch accent (tonally cued prominence) or "*?" for a possible prominence (i.e., uncertainty about presence of a pitch accent). We mapped the pitch accent labels on syllables to words for training a word-level pitch accent classifier with two classes, *accent* and *none*.

Table 1 summarizes the classification performance of the maximum entropy scheme. On a 29k utterance SWBD-DAMSL test set, the dialog act tagger achieves accuracies of 70.4% and 82.9% for the 42 tag and 7 tag vocabulary, respectively. For testing the prosodic prominence classifier, we used a 10-fold cross-validation due to the limited data. Our tagger achieves an accuracy of 78.5% for binary classification of pitch accents (chance = 67.48%). The pitch accent detection accuracy reported here is close to the state-of-the-art for spontaneous speech from the Switchboard corpus (Harper et al., 2001). More details about the automatic prosody labeler can and the dialog act tagger can be found in Rangarajan Sridhar et al. (2008c) and Rangarajan Sridhar et al. (2009), respectively (Table 2).

Table 1
Prosodic prominence and dialog act tagging accuracies for various cues.

Cues used	Accuracy (%)		
	Prominence	DA tags	
		42 tags	7 tags
Lexical	72.6	69.7	81.9
Lexical + syntactic	75.9	70.0	82.4
Lexical + syntactic + prosodic	78.5	70.4	82.9

Table 2
Statistics of the training and test data used in the experiments.

	Training						Test					
	Farsi	Eng	Jap	Eng	Chinese	Eng	Farsi	Eng	Jap	Eng	Chinese	Eng
Sentences	8066		12,239		46,311		925		604		506	
Running words	76,321	86,756	64,096	77,959	351,060	376,615	5442	6073	4619	6028	3826	3897
Vocabulary	6140	3908	4271	2079	11,178	11,232	1487	1103	926	567	931	898
Singletons	2819	1508	2749	1156	4348	4866	903	573	638	316	600	931

5. Data

We report enriched translation experiments on three different parallel corpora: Farsi-English, Japanese-English and Chinese-English. The Farsi-English data used in this paper was collected for doctor–patient mediated interactions in which an English speaking doctor interacts with a Persian speaking patient (Narayanan et al., 2006). The corpus consists of 9315 parallel sentences. The subset was chosen such that each of the English source sentences had corresponding audio. The conversations are spontaneous and the audio was recorded through a microphone (22.5 kHz).

The Japanese-English parallel corpus is a part of the "How May I Help You" (HMIHY) (Gorin et al., 1997) corpus of operator-customer conversations related to telephone services. The corpus consists of 12,239 parallel sentences with corresponding English side audio. The conversations are spontaneous even though the domain is limited. The audio was recorded over a telephone channel (8 kHz). The Chinese-English corpus corresponds to the IWSLT06 training and 2005 development set comprising 46k and 506 sentences, respectively (Paul, 2006). The data are traveler task expressions and are accompanied with Chinese audio. Hence, for DA tag prediction on the English side we used only the lexical and syntactic cues in our Maxent DA tagger. We reserved 5% of the training data, chosen randomly, for development purposes.

6. Exploiting contextual information in S2S translation

In this section, we demonstrate specific techniques to exploit source and target side contextual information in machine-mediated S2S translation. Our treatment is general, i.e., the techniques described here can be used to incorporate other rich contextual information such as semantic concepts, emphasis, emotion and contrast. The only pre-requisite for incorporating other types of contextual information is a reliable tagger that can predict these representations with reasonably high accuracy. In this section, we present methodologies for enriching S2S translation from the source and target side. We categorize contextual information to be of two types, word-level and utterance level. A high-level block diagram of our enriched S2S approach is illustrated in Fig. 2.

6.1. Source language enrichment

Source language enrichment pertains to exploiting contextual information from the source speech or text. Typically, machine translation frameworks exploit only the word or phrase-level information in the translation process. Recently, increasing attention has been given to the use of morphological and syntactic information in machine translation. The introduction of factored translation models has facilitated the use of such multiple knowledge sources at the

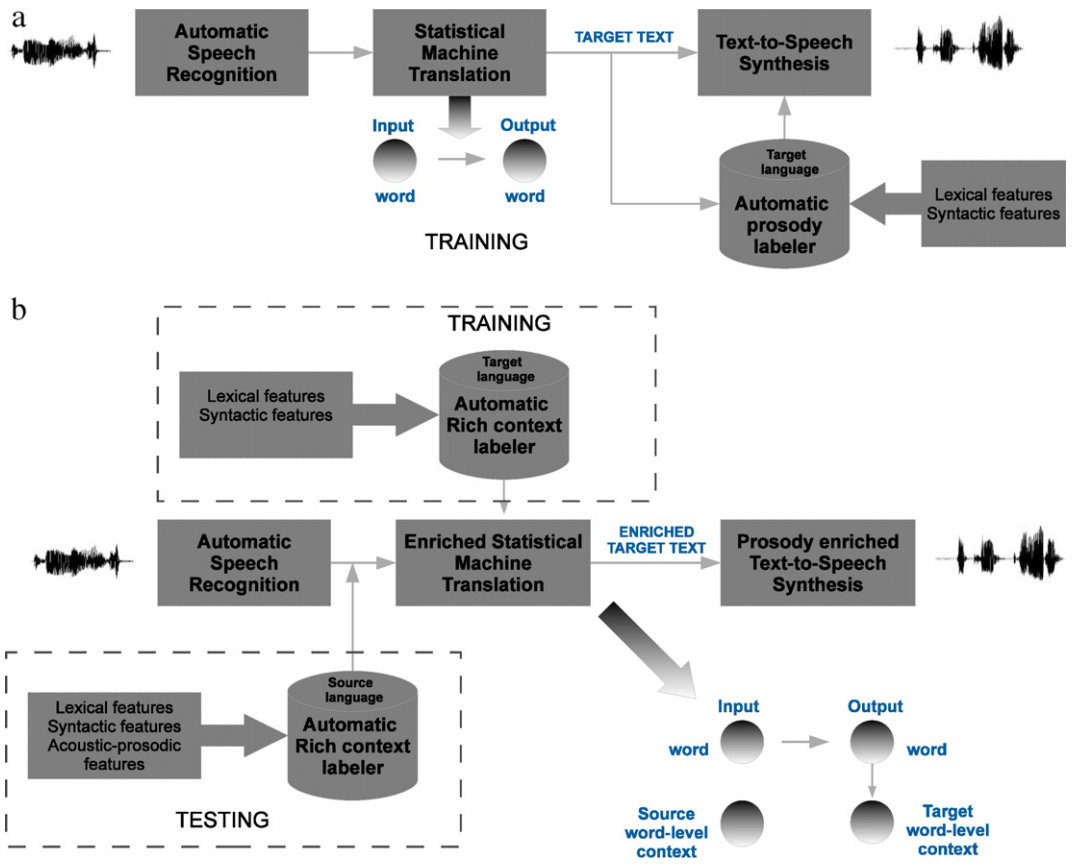

Fig. 2. Illustration of the proposed scheme in comparison with conventional approaches.

word-level without exacerbating the data sparsity problems associated with compound tokens. On the other hand, valuable utterance-level information can also be gleaned from the source speech and text (reference or output of speech recognition). For example, detecting the source utterance's communicative act, emotion, semantic concept, etc., can offer robustness in cross-lingual communication. Such information can help in disambiguation and improved understanding in the target language. At the same time, they can inform the text-to-speech synthesis component to synthesize target speech that carries similar intonation, emotion as the source speech. In our framework, we generate these rich tokens during the testing phase by using pre-trained maximum entropy classifiers.

Suppose, the machine translation component had access to source language contextual information (L_s), the translation problem may be reformulated as

$$T_t^* = \arg\max_{T_t} P(T_t|T_s, L_s) \tag{8}$$

$$\begin{aligned} &= \arg\max_{T_t} \frac{P(T_s|T_t, L_s).P(T_t|L_s)}{P(T_s|L_s)} \\ &= \arg\max_{T_t} P(T_s|T_t, L_s).P(T_t|L_s) \end{aligned} \tag{9}$$

If L_s represents word-level context, Eq. (8) can be used to predict the target text given the source representation and text as in factored translation models (Koehn and Hoang, 2007). On the other hand, if L_s represents utterance-level context, Eq. (9) can be used to split the problem into constructing a context-specific translation model and language model. Hence, depending on the nature of the contextual labels, one can exploit them within the S2S translation paradigm. In this section we demonstrate how source language dialog act tags can be directly exploited in statistical speech translation. We present two speech translation frameworks for exploiting DA tags. The first is a standard phrase

based statistical translation system (Koehn, 2004) and the second is a global lexical selection and reordering approach based on translating the source utterance into a bag-of-words (Bangalore et al., 2007).

6.1.1. Phrase-based translation with dialog act tags

One of the currently popular and predominant schemes for statistical translation is the phrase-based approach (Koehn, 2004). Typical phrase-based SMT approaches obtain word-level alignments from a bilingual corpus using tools such as GIZA++(Och and Ney, 2003) and extract phrase translation pairs from the bilingual word alignment using heuristics. Suppose, the SMT had access to source language dialog act tags (L_s^*), the translation problem may be reformulated as stated in Eq. (9).

The first term in Eq. (9) corresponds to a dialog act specific MT model and the second term to a dialog act tag specific language model. Given sufficient amount of training data such a system can possibly generate hypotheses that are more accurate than the scheme without the use of dialog act tags. However, for small scale and limited domain applications, Eq. (9) leads to an implicit partitioning of the data corpus and might generate inferior translations in terms of lexical selection accuracy or BLEU score. A natural step to overcome the sparsity issue is to employ an appropriate back-off mechanism that would exploit the phrase translation pairs derived from the complete data. A typical phrase translation table consists of 5 phrase translation scores for each pair of phrases, source-to-target phrase translation probability (λ_1), target-to-source phrase translation probability (λ_2), source-to-target lexical weight (λ_3), target-to-word lexical weight (λ_4) and phrase penalty ($\lambda_5 = 2.718$). The lexical weights are the product of word translation probabilities obtained from the word alignments. To each phrase translation table belonging to a particular DA-specific translation model, we append those entries from the baseline model that are not present in phrase table of the DA-specific translation model. The appended entries are weighted by a factor α.

$$\begin{aligned}(T_s \rightarrow T_t)_{L_s^*} &= (T_s \rightarrow T_t)_{L_s} \cup \{\alpha.(T_s \rightarrow T_t) \\ &\quad s.t.\ (T_s \rightarrow T_t) \notin (T_s \rightarrow T_t)_{L_s}\}\end{aligned} \tag{10}$$

where $(T_s \rightarrow T_t)$ is a short-hand[1] notation for a phrase translation table. $(T_s \rightarrow T_t)_{L_s}$ is the DA-specific phrase translation table, $(T_s \rightarrow T_t)$ is the phrase translation table constructed from entire data and $(T_s \rightarrow T_t)_{L_s^*}$ is the newly interpolated phrase translation table. The interpolation factor α is used to weight each of the four translation scores (phrase translation and lexical probabilities for the bilanguage) with the phrase penalty remaining a constant. Such a scheme ensures that phrase translation pairs belonging to a specific DA model are weighted higher and also ensures better coverage than a partitioned data set.

6.1.2. Bag-of-words lexical choice and permutation reordering model

Conventional phrase-based translation relies on learning phrasal associations that are derived from word alignment information. The target bag-of-phrases is typically reordered using a target language model. As a result, there is little emphasis on global lexical reordering which may be necessary for certain language pairs. In contrast, a bag-of-words approach to translation estimates the probability of each target word independently in the context of the entire source sentence. The detected bag-of-words can then be reordered using a language model. Such a bag-of-words (BOW) approach to translation was first presented in Bangalore et al. (2007) and is illustrated in Fig. 3.

In this work, we extend the bag-of-words approach by exploiting dialog act tags and thus, enriching translation. We treat the target sentence as BOWs assigned to the source sentence and its corresponding dialog act tag. The objective here is, given a source sentence and the dialog act tag, to estimate the probability of finding a given word in the target sentence. Since each word in the target vocabulary is detected independently, one can use simple binary static classifiers. The classifier is trained with word n-grams and the dialog act obtained is from the source sentence T_s ($BOW\ grams(T_s)$, L_s). During decoding, the words with conditional probability greater than a threshold θ are considered as the result of lexical choice decoding (Eq. (11)). We use a binary maximum entropy technique with L1-regularization for training

[1] $(T_s \rightarrow T_t)$ represents the mapping between source alphabet sequences to target alphabet sequences, where every pair $(t_1^s, \cdots, t_n^s, t_1^t, \cdots, t_m^t)$ has a weight sequence $\lambda_1, \cdots, \lambda_5$ (five weights).

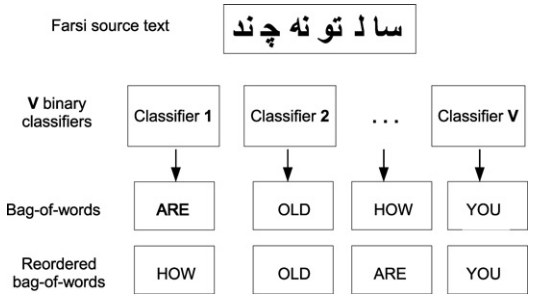

Fig. 3. Illustration of bag-of-words approach to translation.

the bag-of-words lexical choice model. The machine learning toolkit LLAMA (Haffner, 2006) was used for training the Maxent model:

$$BOW^*_{T_t} = \{T_t | P(T_t | BOW\ grams(T_s), L_s) > \theta\} \tag{11}$$

For reconstructing the correct order of words in the target sentence, Bangalore et al. (2007) consider all permutations of words in $BOW^*_{T_t}$ and rank them using a target language model. In this work, we used a separate language model for each dialog act tag, created by interpolating the DA-specific language model with the baseline language model obtained from the entire data. We control the length of the target sentences by varying the parameter θ.

6.2. Target language enrichment

Target language enrichment pertains to exploiting contextual information in the target language. Contextual information in the target language such as prosodic word prominence, contrast and semantic category can be exploited in machine-mediated S2S translation in two ways.

- Post processing: the target contextual labels are produced at the output of the translation block using lexical and syntactic cues from hypothesized text.
- Factored models: the target contextual labels are generated as part of the translation output. Essentially, the translation process translates input source text into output compound tokens that are modeled using a factored approach.

Both the above-mentioned approaches can aid in improving the text-to-speech synthesis component by providing additional linguistic information. However, there is a subtle difference in the incorporation of the enriched target labels. While the target labels for the post-processing approach are predicted in the testing phase using a maximum entropy classifier, the factored approach utilizes target labels that are predicted from the training data. During the testing phase, the factored model generates the target text along with the associated contextual labels.

Suppose, the machine translation component had access to target language word-level contextual information (L_t), the translation problem may be reformulated as

$$\max_{S_t} P(S_t|S_s) \approx \max_{S_t} P(S_t|T^*_t, L^*_t) \times \max_{T_t, L_t} P(T_t, L_t|T^*_s) \times \max_{T_s} P(T_s|S_s) \tag{12}$$

where T^*_s is the output of speech recognition, T^*_t and L^*_t are the target text and enriched contextual representation obtained from translation. While conventional approaches address the detection of L^*_t separately through postprocessing, we integrate this within the translation process. The rich annotations (L_t) can be syntactic, semantic concepts (Koehn and Hoang, 2007; Gu et al., 2006), prosodic information, etc.

In this work, we incorporate prosodic prominence (represented through categorical *pitch accent* labels) in a statistical speech translation framework by injecting these labels into the target side of translation. Our approach generates enriched tokens on the target side in contrast to conventional systems that predict prosody from the output of the statistical machine translation using just hypothesized text and syntax. The proposed framework integrates the assignment of prominence to word tokens within the translation engine. Furthermore, the enriched target language output can

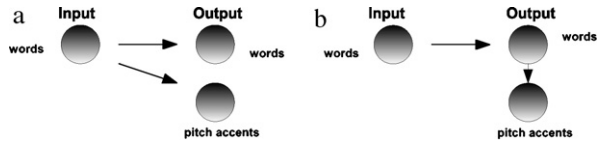

Fig. 4. Illustration of the proposed factored translation models to incorporate prominence. (a) Factored model 1 and (b) factored model 2.

be used to facilitate prosody enriched text-to-speech synthesis, the quality of which is typically preferred by human listeners (Strom et al., 2007).

6.2.1. Factored translation models for incorporating prosodic word prominence

Factored translation models (Koehn and Hoang, 2007) have been proposed recently to integrate linguistic information such as part-of-speech, morphology and shallow syntax in conventional phrase-based statistical translation. The framework allows for integrating multiple levels of information into the translation process instead of incorporating linguistic markers in either preprocessing or postprocessing. For example, in morphologically rich languages it may be preferable to translate lemma, part-of-speech and morphological information separately and combine the information on the target side to generate the output surface words. While these models can be used to incorporate both source and target side word-level contextual information, we focus on models of latter type.

Factored translation models have been used primarily to improve the word level translation accuracy by incorporating the factors in phrase-based translation. In contrast, we are interested in integrating factors such as pitch accent labels in speech translation with the objective of maximizing the accuracy of the output factors themselves. By facilitating factored translation with pitch accent labels predicted from prosodic, syntactic and lexical cues, our enriched translation scheme can produce output with improved pitch accent assignment accuracy. On the other hand, predicting prominence at output of conventional S2S systems is subject to greater error due to typically noisy translations and lack of direct acoustic-prosodic information. Fig. 4 illustrates the type of factored models used in this work.

Factored model 1 represents joint translation of words and prominence. Thus, the phrase translation table obtained for such a model would have compound tokens (word + prominence) in the target language. However, with a factored approach we can build the alignments based on the words alone, thus avoiding data sparsity typically introduced by compound tokens. Factored model 2 translates input words to output words and generates prominence labels from the output word forms through a generation step.

7. Experimental results

In this section, we present experimental results for both source side and target side enrichment using dialog act tags and prosodic prominence, respectively. The phrase-based translation experiments reported in this work were conducted using the Moses[2] toolkit (Koehn et al., 2007) for statistical machine translation. Training the translation models starts from the parallel sentences from which we learn word alignments by using GIZA++toolkit (Och and Ney, 2003). The bidirectional word alignments obtained using GIZA++were consolidated by using the *grow-diag-final* option in Moses. Subsequently, we learn phrases (maximum length of 7) from the consolidated word alignments. A lexicalized reordering model (*msd-bidirectional-fe* option in Moses) was used for reordering the phrases in addition to the standard distance based reordering (*distortion-limit* of 6). The language models were interpolated Kneser–Ney discounted trigram models, all constructed using the SRILM toolkit. Minimum error rate training (MERT) was performed on a development set to optimize the feature weights of the log-linear model used in translation. During decoding, the unknown words were dropped from the hypotheses.

7.1. Incorporating dialog act tags in S2S translation

In all our experiments we assume that the same dialog act is shared by a parallel sentence pair. Thus, even though the dialog act prediction is performed for English target, we use the predicted dialog act as the dialog act for the source

[2] http://www.statmt.org/moses.

Table 3
F-measure and BLEU scores for the two different translation schemes with and without use of dialog act tags. All BLEU score improvements except Japanese-English are statistically significant at $p = 0.05$.

Framework	Language pair	*F*-score (%)			BLEU (%)		
		w/o DA tags	w/ DA tags		w/o DA tags	w/ DA tags	
			7 tags	42 tags		7 tags	42 tags
BOW model	Farsi-Eng	58.00	59.14	59.35	15.95	16.99	17.12
	Japanese-Eng	79.50	79.82	79.93	42.54	44.70	44.98
	Chinese-Eng	68.83	69.70	69.91	54.76	55.98	56.14
Phrase-based translation	Farsi-Eng	56.46	57.32	57.74	22.90	23.50	23.75
	Japanese-Eng	79.05	79.40	79.51	54.15	54.21	54.32
	Chinese-Eng	65.85	67.24	67.49	48.59	52.12	53.04

language sentence. For the test sentences, the sentences in the source language were first translated into English using the model trained on the training data (and optimized using MERT) and subsequently, the tagger was used on the translated text.

The phrase table interpolation weight (α) was obtained by performing a greedy search on the development set. The value of α was set to 0.01. The lexical selection accuracy and BLEU scores for the three parallel corpora using the two schemes described in Section 6.1.1 and 6.1.2 are presented in Table 3. Lexical selection accuracy is measured in terms of the F-measure derived from recall ($|Res \cap Ref|/|Ref| \times 100$) and precision ($|Res \cap Ref|/|Res| \times 100$), where *Ref* is the set of words in the reference translation and *Res* is the set of words in the translation output.

For both the statistical translation frameworks, adding dialog act tags (either 7 or 42 tag vocabulary) consistently improves both the lexical selection accuracy and BLEU score for all the language pairs. The improvements in BLEU score are 95% statistically significant (Koehn, 2004) for Farsi-English and Chinese-English but not for Japanese-English. This can be attributed to the distribution of dialog act tags in the Japanese-English corpus. Almost 90% of the corpus comprises of *statements* and hence the phrase table interpolation does not provide any significant improvement over the baseline table constructed from the entire training data. While the BOW model provides higher lexical selection accuracy, the phrase-based translation provides better BLEU score. In the BOW model, we detect each word in the target vocabulary independently and reorder the bag-of-words separately. The framework focuses on maximizing the occurrence of target words in the context of a given source sentence. Further, the permutation model used for reordering is still inferior to state-of-the-art reordering techniques. Hence, the lexical selection accuracy reported in this work is higher in comparison to the BLEU score. On the other hand, phrase-based translation produces a bag-of-phrases in the target language which are reordered using a distortion model. The framework focuses on maximizing the occurrence of target phrases in the context of source phrases and can potentially generate target hypotheses with both high lexical selection accuracy and BLEU score (weighted *n*-gram precision).

Next, we investigate the contribution of each dialog act to the overall improvement in translation quality. We analyze the performance in terms of lexical selection accuracy and BLEU score improvements per dialog act. Fig. 5 shows the distribution of dialog acts in the 7 vocabulary dialog act tag set across the three corpora used in our experiments. *Statements* are the most frequent dialog acts followed by *question*, *other* and *acknowledgment*. Dialog acts such as *agreement, appreciation* and *abandoned* occur quite infrequently in the corpora.

In Table 4, we report the lexical selection accuracies and BLEU scores per dialog act for the BOW model and phrase-based translation model, respectively, on the Farsi-English corpus. The table compares the per DA performance of the two translation models with and without the use of dialog act information in the translation process. The BLEU scores were computed on subsets of sentences in the test corpus that belonged to a particular category of dialog act tags. The BLEU score improvements for *statements*, *questions*, *other* and *acknowledgments* are statistically significant at $p = 0.05$ while the improvement in other categories is not statistically significant due to the small number of samples. Since, *acknowledgments* are typically short sentences with a constrained vocabulary, they result in higher BLEU scores in comparison with *statements* and *questions* that are more open-ended. The results indicate that knowledge of discourse context such as *question* or *acknowledgment* is most beneficial to the translation process. Knowledge of detecting an utterance as a *statement* does not offer any significant improvement in the translation. This may be

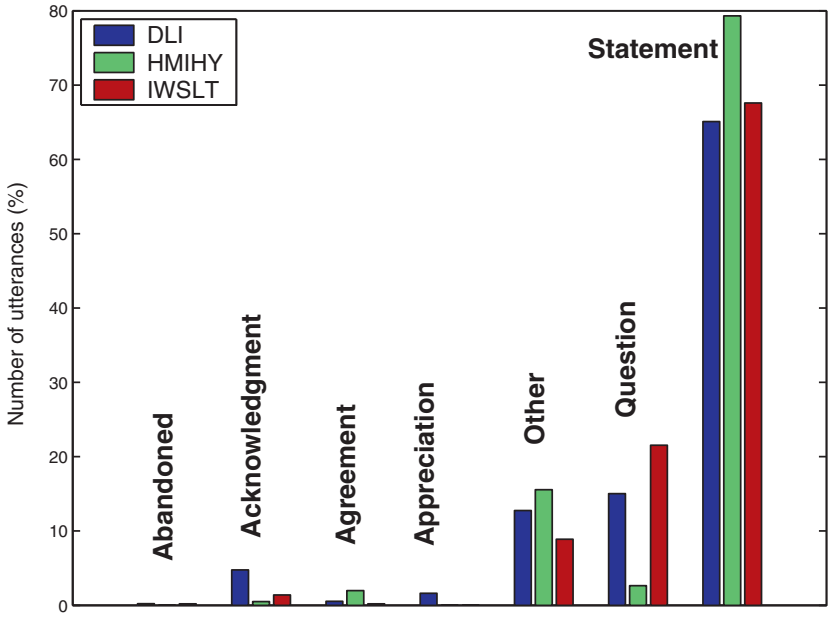

Fig. 5. Distribution of dialog acts in the test data of each corpus.

Table 4
Lexical selection accuracy (%) and BLEU score (%) per DA tag for the BOW model and phrase-based translation scheme with and without use of dialog act tags for the DLI Farsi-English corpus. The BLEU score improvements for *Statement, Question, Other, Acknowledgment* are statistically significant at $p = 0.05$.

Dialog act	BOW model Lexical accuracy		Phrase-based BLEU	
	w/o DA	w/ DA	w/o DA	w/ DA
Statement	55.82	56.31	20.58	20.57
Question	59.85	62.14	24.12	26.36
Other	71.05	69.09	37.84	41.19
Acknowledgment	85.22	87.04	51.21	69.30
Appreciation	71.05	76.32	46.92	73.02
Agreement	56.00	66.67	18.46	50.00
Abandoned	75.00	75.00	58.41	58.41

attributed to lack of systematic structural information (syntactic) or cue words that differentiate *statements* from other dialog acts. Deeper analysis using the 42 DA tag set indicates that dialog acts such as *yes–no questions*, *wh-questions* and *open questions* contribute the most to the lexical selection accuracy and BLEU score improvement. Similar trends hold for the Chinese-English corpus. On the other hand, the improvements for the Japanese-English corpus is largely insignificant due to the high proportion of *statements* in the test corpus.

7.2. Incorporating prosodic prominence in S2S translation

We report results for three scenarios which vary in terms of how the prominence labels are produced in the target language.

1 Post processing: the pitch accent labels are produced at the output of the translation block using lexical and syntactic cues from hypothesized text.
2 Factored model 1: factored model that translates source words to target words and pitch accents.
3 Factored model 2: factored model that translates source words to target words which in turn generate pitch accents

Table 5
Evaluation metrics for the two corpora used in experiments (all scores are in %).

Translation model	Farsi-English			Japanese-English		
	Lexical *F*-score	BLEU	Prosodic accuracy	Lexical *F*-score	BLEU	Prosodic accuracy
Postprocessing	56.46	22.90	74.51	78.98	54.01	68.57
Factored model 1	56.18	22.86	80.83	79.00	53.98	80.12
Factored model 2	56.07	22.75	80.57	78.56	53.87	79.56

Source : من من برای آسم دارو مصرف میکنم
Reference : I_none I'm_* taking_none medication_* for_none asthma_*
Hypothesis: I'm_none on_none medication_* for_none asthma_*

Ref ∩ Res : {I'm, medication, for, asthma}
#correct pitch accents: 3

Fig. 6. Illustration of the process used to calculate prosodic accuracy.

Table 5 summarizes the results obtained in terms of BLEU score (Papineni et al., 2002), lexical selection accuracy and prosodic accuracy. Lexical selection accuracy is measured in terms of the F-measure derived from recall ($|Res \cap Ref|/|Ref| \times 100$) and precision ($|Res \cap Ref|/|Res| \times 100$), where *Ref* is the set of words in the reference translation and *Res* is the set of words in the translation output. Prosodic accuracy is defined as $\sharp\, correct\ pitch\ accents \in (Res \cap Ref)/|Res \cap Ref| \times 100$. Fig. 6 illustrates the computation of prosodic accuracy for an example utterance.

The reference pitch accent labels for the English sentences were obtained from the automatic prominence labeler described in Section 4 using lexical, syntactic and prosodic cues. The language models were trigram models created only from the training portion of each corpus. The results in Table 5 indicate that the assignment of correct pitch accents to the target words improves with the use of factored translation models. Factored model 1 that translates input word forms to output word forms and pitch accents achieves the best performance. We obtain a relative improvement of 8.4% and 16.8% in prosodic accuracy for the two corpora in comparison with the postprocessing approach. In the postprocessing approach, the pitch accent classifier was trained on lexical, syntactic and acoustic-prosodic features from clean sentences, but evaluated on possibly erroneous machine translation output. Furthermore, the lack of acoustic-prosodic information at the output of machine translation results in lower prosodic assignment accuracy. On the other hand, factored models integrate the pitch accent labels derived from lexical, syntactic and acoustic-prosodic features within the translation framework. Thus, the prosodic accuracy obtained is consistently higher than the postprocessing scheme.

Table 5 also illustrates translation performance at the word level. For both the factored translation models, the word-level BLEU score and lexical selection accuracy are close to the baseline model that uses no pitch accent labels within the translation framework. Thus, the improvement in prosodic assignment accuracy is obtained at no significant degradation of the word-level translation performance.[3]

8. Discussion

It is important to note that the dialog act tags and prosodic prominence labels used in our translation system are predictions from the Maxent based DA tagger described in Section 4. We do not have access to the reference tags; thus, some amount of error is to be expected in the DA and prosody tagging. Despite the lack of reference DA tags, we are still able to achieve modest improvements in the translation quality. Improving the current DA tagger and developing suitable adaptation techniques are part of future work.

The results reported in Table 1 demonstrate improvements in BLEU score by exploiting dialog act tags in the translation process. However, the results do not separate the impact of using dialog act tags in constructing the phrase

[3] The BLEU score for Japanese-English is different from the baseline score in Table 3 as two sentences were removed from the test set due to audio clipping issues that the prosody tagger could not process.

Table 6
BLEU score comparison using the dialog act tags in phrase table only and using it in both phrase table and language model of the phrase-based translation model (DA tagger with 7 tags was used in the experiment).

Phrase-based Translation Model	BLEU(%)		
	Farsi-English	Japanese-English	Chinese-English
No DA tags	22.90	54.15	48.59
DA interpolated phrase table	23.44	54.23	51.94
DA interpolated phrase table and language model	23.50	54.21	52.12

translation table and language model separately. In order to better understand the accuracy improvement stemming from the use of dialog act tags, we performed translation with and without the use of dialog act information in the language model. The language model unaware of dialog act information was simply constructed from the entire training data while the dialog act aware language model was built by interpolating the entire training data with dialog act specific sentences from the training data. The interpolation weights were learned by minimizing the language model perplexity on a development set. The results are shown in Table 6. The results show that using dialog act information in the phrase table interpolation by itself offers significant improvement over not using any dialog act information. The use of dialog act specific language model marginally improves the BLEU score but the improvement is not statistically significant.

The phrase translation interpolation described in Section 6.1.1 uses the same interpolation factor (α) for each of the four translation model weights. Further, the same factor is used across all the dialog acts. Optimizing each of the translation model weights separately and independently for each dialog act subset of the data can possibly generate better translations. The dialog act specific translation models use the same feature weights as that of the baseline model optimized using MERT. Using a large development set that comprises different dialog types can be used to optimize the feature weights for each dialog act model separately and is likely to provide improved accuracy. The work described here is better suited for translation scenarios that do not involve multiple sentences as part of a turn (e.g., lectures or parliamentary addresses). However, this is not a principled limitation of the proposed work. It can be overcome by using the DA tagger on segmented utterances (or sentences separated by punctuation). Furthermore, the experiments in this paper have been performed on reference transcripts. We plan to evaluate our framework on speech recognition output as well as lattices as part of our future work.

While we have demonstrated that our framework can improve the accuracy of prominence labels in the target language, it can potentially be used to integrate any word-level rich annotation dependent on prosodic features (e.g., boundary tones, emotion, etc.). For example, the factored translation models proposed in this work may be especially useful for tonal languages such as Chinese where it is important to associate accurate tones to syllables. The proposed framework needs to be evaluated by including a speech synthesis system that can make use of prosodic markers. We plan to address this also as part of future work. Finally, while we have demonstrated that using rich contextual information can improve translation quality in terms of word based automatic evaluation metrics, the real benefits of such a scheme would be manifested through human evaluations. We are currently working on conducting subjective evaluations.

9. Challenges and future work

In the following section we outline some of the challenges in exploiting rich contextual features in machine-mediated speech-to-speech translation. We believe that the following research problems are critical to improving the current state-of-the-art in S2S translation.

1 Designing appropriate user interfaces that can augment the translation hypotheses with contextual information.
2 Detecting and exploiting source language prosody and dialog information in target text-to-speech synthesis.
3 Actively learning and adapting the system from user feedback or user in the loop.

In this work, we demonstrated the utility of contextual information in improving objective metrics such as BLEU score and lexical selection accuracy. On the other hand, directly displaying enriched information in machine-mediated

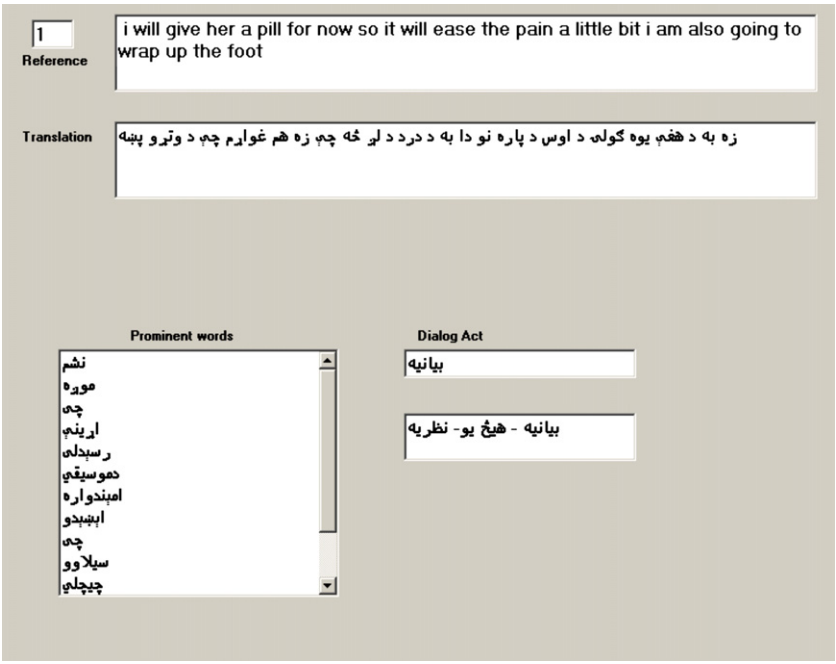

Fig. 7. Example of a user interface that displays contextual information in machine-mediated speech translation (English-Pashto).

S2S translation can aid the end users in better understanding and disambiguation. Furthermore, augmenting the translation hypotheses with enriched information can also aid the users in rephrasing within a dialog. For example, one can display the prominent words and dialog act tags for a particular source sentence and provide the target language user with multiple sources of information to appropriately decode and understand the input sentence in context. Fig. 7 illustrates a sample user interface that displays enriched information in addition to the translation hypothesis. We believe that augmenting conventional user interfaces with rich information such as dialog act tags, prominent words, topics, named entities, semantic concepts can greatly improve the end-to-end communication as opposed to objective metrics that simply measure phrase-level accuracy.

The second aspect that needs attention is the appropriate use of rich information in target language text-to-speech synthesis. Even though text-to-speech synthesis has advanced significantly and can produce natural sounding speech with appropriate intonation, S2S systems have to work with translation hypotheses that are often inaccurate. Moreover, current S2S systems typically perform text-to-speech synthesis using models trained on independent corpora that completely ignore the source language context. For example, a source sentence that is a *question* can be synthesized in the target language as a statement due to translation errors. The framework presented in this work can inform the target text-to-speech synthesis with source language context such as dialog act, emotion, emphasis and contrast. Incorporating such contextual information is a key challenge in improving the overall quality of S2S translation. Recently, Rangarajan Sridhar et al. (2011) have shown that the use of automatic dialog act tags (generated by the tagger described in this work) can significantly improve text-to-speech synthesis quality through subjective evaluation tests.

Finally, most S2S translation systems are optimized for objective evaluation metrics. Little importance has been given to exploiting user feedback or user in the loop for active learning (Cohn et al., 1996). For example, users can be provided with multiple translation hypotheses and prompted to indicate the most preferred translation. The preferred bilingual sentence pairs can either be used for retraining the translation model or the constituent bilingual phrase pairs can be weighted more for future runs. One can also use the user in the loop to refine rich information. The prominent words box shown in Fig. 7 can show the corresponding translation of each word which can then be accepted or rejected by the user. The users can also choose to provide information regarding dialog state, semantic tags or named entities that can be used in subsequent enrichment. These are just a few examples of incorporating user knowledge into a translation system. In order to have tractable S2S translation systems that respond to the individual needs of a user, it is critical to exploit user feedback for adaptation and active learning.

10. Conclusions

Conventional approaches to speech-to-speech translation typically ignore key contextual information such as prosody, emphasis, discourse state in the translation process. In this work, we demonstrated how S2S translation systems can benefit from exploiting rich information, particularly, machine-mediated cross-lingual translation systems. We formulated an enriched S2S translation framework in contrast to the traditional pipeline architecture that ignores contextual information. The framework presented here can exploit source side and target side rich representations either at the word or utterance level. The only prerequisite to exploiting the rich representations is the availability of automatic classifiers that can reliably detect the contextual information.

First, we presented a maximum entropy framework for automatic classification of dialog act tags and prosodic word prominence. Subsequently, we demonstrated techniques for contextual enrichment from the source and target side. We focused on integrating dialog act tags from the source side of translation and prosodic prominence on the target side of the translation. The dialog act tags provided significant improvements in BLEU score and lexical selection accuracy on three different parallel corpora. While the phrase table interpolation approach was found to improve the BLEU score, the BOW model was more found to be more suitable for obtaining higher lexical selection accuracy. We also found that finer representation of dialog information can be more beneficial in improving translation accuracy in comparison with coarse grained DA tags. Specifically, *yes-no questions*, *wh-questions* and *open questions* contribute the most to the lexical selection accuracy and BLEU score improvement.

We also demonstrated how prosodic word prominence can be exploited on the target side of S2S translation using factored translation models. We presented two types of factored translation models for generating enriched word tokens and compared them with the conventional approach of using post-processing labeling. Integrating the prosodic prominence labeling within the translation process provided significant improvement in pitch accent labeling accuracy in comparison with a post-processing approach. In principle, one can use other types of contextual tags such as emotion, affect and contrast using the framework presented in this work. We are currently working on conducting subjective evaluations to ratify the usefulness of augmented information for understanding and disambiguation. We are also working on exploiting the augmented information in target language text-to-speech synthesis.

References

Agüero, P.D., Adell, J., Bonafonte, A., 2006. Prosody generation for speech-to-speech translation. In: Proceedings of ICASSP, Toulouse, France, May.

Avaramidis, E., Koehn, P., 2008. Enriching morphologically poor languages for statistical machine translation. In: Proceedings of ACL.

Bangalore, S., Joshi, A.K., 1999. Supertagging: an approach to almost parsing. Computational Linguistics 25 (June (2)).

Bangalore, S., Haffner, P., Kanthak, S., 2007. Statistical machine translation through global lexical selection and sentence reconstruction. In: Proceedings of ACL.

Bennett, W.S., 1989. The place of semantics in MT systems. Literary and Linguist Computing 4 (3), 200–202.

Bertoldi, N., Zens, R., Federico, M., Shen, W., 2008. Efficient speech translation through confusion network decoding. IEEE Transactions on Audio, Speech, and Language Processing 16 (November (8)), 1696–1705.

Black, A.W., Hunt, A.J., 1996. Generating F0 contours from ToBI labels using linear regression. In: Proceedings of ICSLP 96, vol. 3, pp. 1385–1388.

Bulyko, I., Ostendorf, M., 2001. Joint prosody prediction and unit selection for concatenative speech synthesis. In: Proc. of ICASSP.

Byron, D., Shriberg, E., Stolcke, A., 2002. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In: Proc. of ICSLP, vol. 2, Denver, pp. 949–952.

Campbell, W.N., Black, A.W., 1996. Prosody and the selection of source units for concatenative synthesis. In: Progress in Speech Synthesis, pp. 279–292.

Cohn, D.A., Ghahramani, Z., Jordan, M.I., 1996. Active learning with statistical models. Journal of Artificial Intelligence Research 4, 129–145.

Dorr, B.J., 1992. The use of lexical semantics in interlingual machine translation. Machine Translation 7 (3), 135–193.

Fügen, C., Kolss, M., 2007. The influence of utterance chunking on machine translation performance. In: Proceedings of Interspeech, Antwerp, Belgium.

Gao, Y., Zhou, B., Gu, L., Sarikaya, R., Kuo, H., Rosti, A.-V., Afify, M., Zhu, W., 2006. IBM Mastor: multilingual automatic speech-to-speech translator. In: Proceedings of ICASSP, May.

Gorin, A., Riccardi, G., Wright, J., 1997. How May I Help You? Speech Communication 23, 113–127.

Gu, L., Gao, Y., Liu, F.H., Picheny, M., 2006. Concept-based speech-to-speech translation using maximum entropy models for statistical natural concept generation. IEEE Transactions on Audio, Speech and Language Processing 14 (March (2)), 377–392.

Haffner, P., 2006. Scaling large margin classifiers for spoken language understanding. Speech Communication 48 (iv), 239–261.

Harper, M., Dorr, B., Roark, B., Hale, H., Shafran, Z., Liu, Y., Lease, M., Snover, M., Young, L., Stewart, R., Krasnyanskaya, A., 2005. Parsing speech and structural event detection. JHU Summer Workshop, Tech. Rep.

Hassan, H., Sima'an, K., Way, A., 2007. Supertagged phrase-based statistical machine translation. In: Proceedings of ACL.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, S., Taylor, P., Van Ess-Dykema, C., 1998. Switchboard Discourse Language Modeling Project Report. Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD (Technical Report Research Note 30).

Katae, N., Kimura, S., 1996. Natural prosody generation for domain specific text-to-speech systems. In: Proc. of ICSLP, vol. 3, October, pp. 1852–1855.

Kei Fujii, H.K., Campbell, N., 2003. Target cost of f0 based on polynomial regression in concatenative speech synthesis. In: Proceedings of ICPhS.

Koehn, P., Hoang, H., 2007. Factored translation models. In: Proceedings of EMNLP.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C.J., Bojar, O., Constantin, A., Herbst, E., 2007. Moses: open source toolkit for statistical machine translation. In: Proceedings of ACL.

Koehn, P., 2004. Pharaoh: a beam search decoder for phrasebased statistical machine translation models. In: Proceedings of AMTA-04, Berlin/Heidelberg, pp. 115–124.

Koehn, P., 2004. Statistical significance tests for machine translation evaluation. In: Proceedings of EMNLP.

Lafferty, J., McCallum, A., Pereira, F., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Procedings of 18th International Conf. on Machine Learning, San Francisco, CA, pp. 282–289.

Lavie, A., Levin, L., Qu, Y., Waibel, A., Gates, D., Gavalada, M., Mayfield, L., Taboada, M., 1996. Dialogue processing in a conversational speech translation system. In: Proc. of ICSLP, October, pp. 554–557.

Levin, A., Gates, D., Lavie, A., Waibel, A., 1998. An interlingua based on domain actions for machine translation of task-oriented dialogues. In: Proc. of ICSLP, pp. 1155–1158.

Liu, Y., Shriberg, E., Stolcke, A., 2003. Automatic disfluency identification in conversational speech using multiple knowledge sources. In: Proc. Eurospeech, Geneva, September, pp. 957–960.

Matusov, E., Kanthak, S., Ney, H., 2005. On the integration of speech recognition and statistical machine translation. In: Proc. of Eurospeech.

Matusov, E., Hillard, D., Magimai-Doss, M., Hakkani-Tür, D., Ostendorf, M., Ney, H., 2007. Improving speech translation with automatic boundary prediction. In: Proceedings of Interspeech, Antwerp, Belgium.

Mayfield, L., Gavalda, M., Ward, W., Waibel, A., 1995. Concept-based speech translation. In: Proc. of ICASSP, May, pp. 97–100.

Nöth, E., Batliner, A., Kießling, A., Kompe, R., Niemann, H., 2000. VERBMOBIL: the use of prosody in the linguistic components of a speech understanding system. IEEE Transactions on Speech and Audio processing 8 (September (5)), 519–532.

Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G., Kawai, H., Jitsuhiro, T., Zhang, J.-S., Yamamoto, H., Sumita, E., Yamamoto, S., 2006. The ATR multilingual speech-to-speech translation system. IEEE Transactions on Audio, Speech, and Language Processing 14 (March (2)), 365–376.

Narayanan, S., Georgiou, P., Sethy, A., Wang, D., Ananthakrishnan, S., Ettelaie, E., Franco, H., Precoda, K., Vergyri, D., Zheng, J., Wang, W., Gadde, R.R., Graciarena, M., Abrash, V., Richey, C., 2006. Speech recognition engineering issues in speech to speech translation system design for low resource languages and domains. In: Proc. of ICASSP, Toulose, France, May.

Ney, H., Och, F.J., Vogel, S., 2000. Statistical translation of spoken dialogues in the verbmobil system. In: Workshop on Multilingual Speech Communication, Kyoto, pp. 69–74.

Och, F.J., Ney, H., 2003. A systematic comparison of various statistical alignment models. Computational Linguistics 29 (1), 19–51.

Ostendorf, M., Shafran, I., Shattuck-Hufnagel, S., Carmichael, L., Byrne, W., 2001. A prosodically labeled database of spontaneous speech. In: ISCA Workshop on Prosody in Speech Recognition and Understanding, pp. 119–121.

Papineni, K., Roukos, S.,Ward, T., Zhu, W.J., 2002. BLEU: a method for automatic evaluation of machine translation, IBM T.J. Watson Research Center, Tech. Rep.

Paul, M., 2006. Overview of the IWSLT 2006 Evaluation Campaign. In: Proc. of the International Workshop on Spoken Language Translation, Kyoto, Japan, pp. 1–15.

Rangarajan Sridhar, V.K., Bangalore, S., Narayanan, S., 2008a. Enriching spoken language translation with dialog acts. In: Proceedings of ACL.

Rangarajan Sridhar, V.K., Bangalore, S., Narayanan, S., 2008b. Factored translation models for enriching spoken language translation. In: Proceedings of InterSpeech.

Rangarajan Sridhar, V.K., Bangalore, S., Narayanan, S., 2008c. Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. IEEE Transactions on Audio, Speech, and Language Processing 16 (May (4)), 797–811.

Rangarajan Sridhar, V.K., Bangalore, S., Narayanan, S., 2009. Combining lexical, syntactic and prosodic cues for improved dialog act tagging. Computer Speech & Language October (23) (4).

Rangarajan Sridhar, V.K., Syrdal, A., Conkie, A., Bangalore, S., 2011. Enriching text-to-speech synthesis using automatic dialog act tags. In: Proceedings of Interspeech.

Raux, A., Black, A., 2003. A unit selection approach to f0 modeling and its application to emphasis. In: IEEE Workshop on Automatic Speech Recognition and Understanding, vol. 3, December, pp. 700–705.

Reithinger, N., Engel, R., 2000. Robust content extraction for translation and dialog processing. In: Wahlster, W. (Ed.), Verbmobil: Foundations of Speech-to-Speech Translation. Springer, pp. 430–439.

Reithinger, N., Engel, R., Kipp, M., Klesen, M., 1996. Predicting dialogue acts for a speech-to-speech translation system. In: Proc. of ICSLP, October, pp. 654–657.

Ross, K., Ostendorf, M., 1999. A dynamical system model for generating fundamental frequency for speech synthesis. IEEE Transactions on Speech and Audio Processing 7 (May (3)), 295–309.

Sakai, S., Glass, J., 2003. Fundamental frequency modeling for corpus-based speech synthesis based on a statistical learning technique. In: Proceedings of IEEE ASRU.

Stallard, D., Choi, F., Kao, C.L., Krstovski, K., Natarajan, P., Prasad, R., Saleem, S., Subramanian, K., 2007. The BBN 2007 displayless English/Iraqi speech-to-speech translation system. In: Proceedings of InterSpeech.

Strom, V., Nenkova, A., Clark, R., Vazquez-Alvarez, Y., Brenier, J., King, S., Jurafsky, D., 2007. Modelling prominence and emphasis improves unit-selection synthesis. In: Proceedings of Interspeech, Antwerp, Belgium.

Syrdal, A.K., Kim, Y.-J., 2008. Dialog speech acts and prosody: considerations for TTS. In: Proceedings of Speech Prosody.

Wightman, C.W., Ostendorf, M., 1994. Automatic labeling of prosodic patterns. IEEE Transactions on Audio and Speech Processing 2 (3), 469–481.

Yamada, K., Knight, K., 2001. A syntax-based statistical translation model. In: Proceedings of the Conference of the Association for Computational Linguistics, New Brunswick, NJ, pp. 132–139.